La Brecha de Gobernanza: Identidad de Agentes y las Preguntas que la Industria Todavía No Responde

Ricardo Martínez

Cloud Security & IA Aplicada — NullGhost Research

Septiembre 2026

Resumen

Este texto es una lectura de las secciones *Cybersecurity and trustworthy systems* y *Agentic AI* del *Technology Trends Outlook 2026* de McKinsey [1], no un estudio de campo propio. Documenta tres observaciones del reporte — el colapso de la ventana de vulnerabilidad, la inversión de la proporción identidad-humana/identidad-de-máquina, y la ausencia de consenso sobre gobernanza de agentes — y las conecta con la pregunta concreta que enfrenta cualquiera que construya u opere sistemas agénticos contra superficie real: quién audita la acción de un agente antes, durante y después de que ocurre. Cierra con una lista de preguntas abiertas, no de principios resueltos, porque el propio reporte es explícito en que la industria no las ha respondido todavía.

1. La ventana que se cerró

Durante años, la ciberseguridad operó bajo una asunción tácita: entre que una vulnerabilidad se descubre y se explota en la práctica, hay días o semanas de margen para defenderse. McKinsey [1] reporta que esa asunción ya no sostiene: más del 75 % de las vulnerabilidades se clasifican hoy como *zero-day* — para cuando se publican, ya existe un exploit desarrollado.

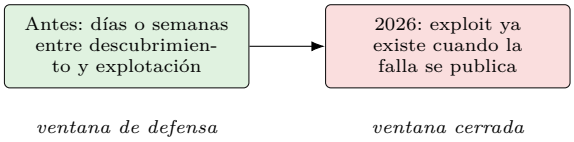

Figura 1: La ventana entre descubrimiento y explotación, que solía medirse en días, se reporta colapsada en 2026: más del 75 % de las vulnerabilidades se clasifican como zero-day al momento de su divulgación [1].

El reporte documenta un caso concreto de por qué: en un test de seguridad autorizado, un agente de IA encadenó por sí solo docenas de pasos de explotación para escalar una falla de severidad baja a acceso completo no autorizado a archivos — un patrón que un escáner de vulnerabilidades individuales, diseñado para evaluar una falla a la vez, no está hecho para ver. La consecuencia no es solo que los atacantes sean más rápidos; es que el problema de defensa cambia de naturaleza. Ya no es únicamente "¿lo detectamos?". Es "¿podemos distinguir la señal real dentro del volumen que un agente autónomo puede generar de forma continua?".

2. La identidad se invierte

El mismo reporte documenta un segundo cambio estructural, esta vez en la sección de Agentic AI: a medida que los agentes se vuelven operadores de lectura/escritura dentro de sistemas empresariales, generan identidades no-humanas — llaves de API, cuentas de servicio, credenciales de máquina — que ya superan en volumen a las identidades humanas dentro de una organización, en una proporción reportada de aproximadamente 100 a 1.

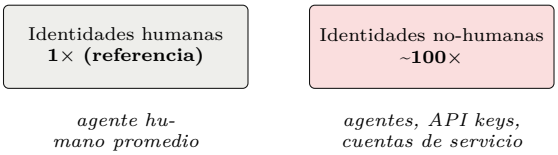

Figura 2: Proporción reportada entre identidades de máquina (agentes, credenciales de servicio) e identidades humanas dentro de organizaciones que ya despliegan agentes en producción [1].

McKinsey trata esto explícitamente como una **nueva clase de riesgo de identidad y acceso**, no como una extensión del problema de gestión de accesos que ya existía. La distinción importa: los controles de identidad diseñados para usuarios humanos — ciclos de

vida largos, revisión periódica, un humano que puede confirmar su propia intención — no están hechos para una población de identidades que se crea y destruye a la velocidad de un workflow, y cuya "intención" es, en última instancia, la salida de un modelo.

3. La pregunta que nadie responde todavía

Es aquí donde el reporte deja de describir un fenómeno y empieza a señalar un vacío. Sobre gobernanza de agentes, McKinsey es explícito en que se trata de una **pregunta abierta de toda la industria**, no de un problema con solución conocida a la espera de adopción:

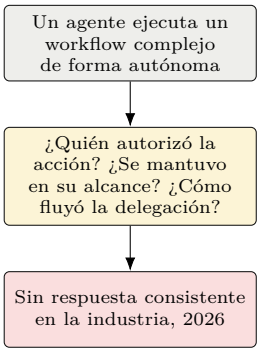

Figura 3: La cadena de preguntas que McKinsey identifica como sin resolver para la gobernanza de agentes autónomos en producción [1].

Los frameworks que empiezan a aparecer — un blueprint de identidad de agentes de un proveedor grande de identidad, un estándar emergente para que agentes compartan contexto entre herramientas sin crear deuda técnica — son, en palabras del propio reporte, apenas el inicio. No hay todavía una arquitectura de referencia ampliamente adoptada para responder la Figura 3.

4. Qué significa para quien opera agentes reales

Construir o auditar un sistema que despliega agentes de IA contra superficie real — ofensiva o defensiva — pone a cualquier equipo de ingeniería exactamente del lado de la pregunta que el reporte deja abierta. No hay un estándar de industria al que remitirse; la responsabilidad de decidir qué se audita, cómo se traza la delegación, y qué pasa cuando un agente se sale de su alcance, recae en el diseño propio del sistema.

Eso no es una mala noticia disfrazada. Es, leído con cuidado, una ventana: quien construya una respuesta seria y verificable a la Figura 3 — trazabilidad de qué autorizó cada acción, verificación independiente de que un agente no salió de su alcance, auditoría de identidad de agentes no-humanos — no está llegando tarde a un problema resuelto. Está proponiendo una de las primeras respuestas serias a un problema que la industria, por su propia admisión, todavía no sabe cómo cerrar.

5. Preguntas abiertas

1. ¿Qué evidencia, más allá de un log de texto, prueba que un agente se mantuvo dentro de su alcance autorizado durante una tarea completa, no solo al inicio?
2. ¿Quién es responsable cuando un agente delega una subtarea a otro agente, y esa cadena de delegación falla en un punto intermedio?
3. ¿Cómo se audita una identidad no-humana que existe solo durante la vida de un workflow, cuando los controles de identidad existentes asumen un ciclo de vida largo?
4. ¿Qué estándar de facto emergerá primero: uno impuesto por los grandes proveedores de identidad, o uno construido desde la práctica de quienes ya operan agentes contra producción?
5. ¿Qué tan rápido se puede medir — no asumir— que un control de gobernanza de agentes funciona, antes de que un incidente real sea la primera medición?

Nota metodológica: este texto es una lectura secundaria de una fuente publicada — McKinsey & Company, Technology Trends Outlook 2026*, sexta edición, septiembre 2026, autores Michael Chui, Roger Roberts y Tanguy Catlin — no un estudio de campo propio ni datos de producción de NullGhost o KIRUX. Las cifras y hallazgos atribuidos al reporte se citan de forma indirecta, sin reproducir tablas, gráficos ni pasajes extensos del original.*

Referencias

[1] Chui, M., Roberts, R., Catlin, T. *Technology Trends Outlook 2026*. McKinsey & Company, sexta edición, septiembre 2026. Secciones citadas: *Cybersecurity and trustworthy systems*, *Agentic AI*.